Linguistic Structure in Aggregate Variation

John Nerbonne
Rijksuniversiteit Groningen

Summer, 2005

Aggregation in Variation

Thesis: Language variation must be studied in the aggregate.

- Detailed studies of single features ([aɪ] vs. [a], [æ] vs. [æ^{ə}]) are at best inconclusive, at worst misleading.
- Bloomfield (1933) noted how confusing details are; Coseriu (^{1}1956,1975) warned against "atomism" in dialectology.
- But question: is the aggregate linguistically structured?

We focus here on the question of linguistic structure.

Outline

- Question
- Aggregating Technique
- Experiment on Southern Vowels in LAMSAS
- Results
- Reflections

Question

Aggregate pronunciation distance:

- Is reliable, given > 20 pronunciations/site (Cronbach $\alpha > 0.8$)
- Correlates with naive speakers' judgements ($r \approx 0.65$) Gooskens & Heeringa (2003), Heeringa (2004: Chap. 7)
- Is predictable from geography (Heeringa & Nerbonne, 2001)
- Provides analytic foundation for dialect continua as organizing principle

But there's little assumption of linguistic structure in this work.

Question: What linguistic elements determine aggregate pronunciation distance (if any)?

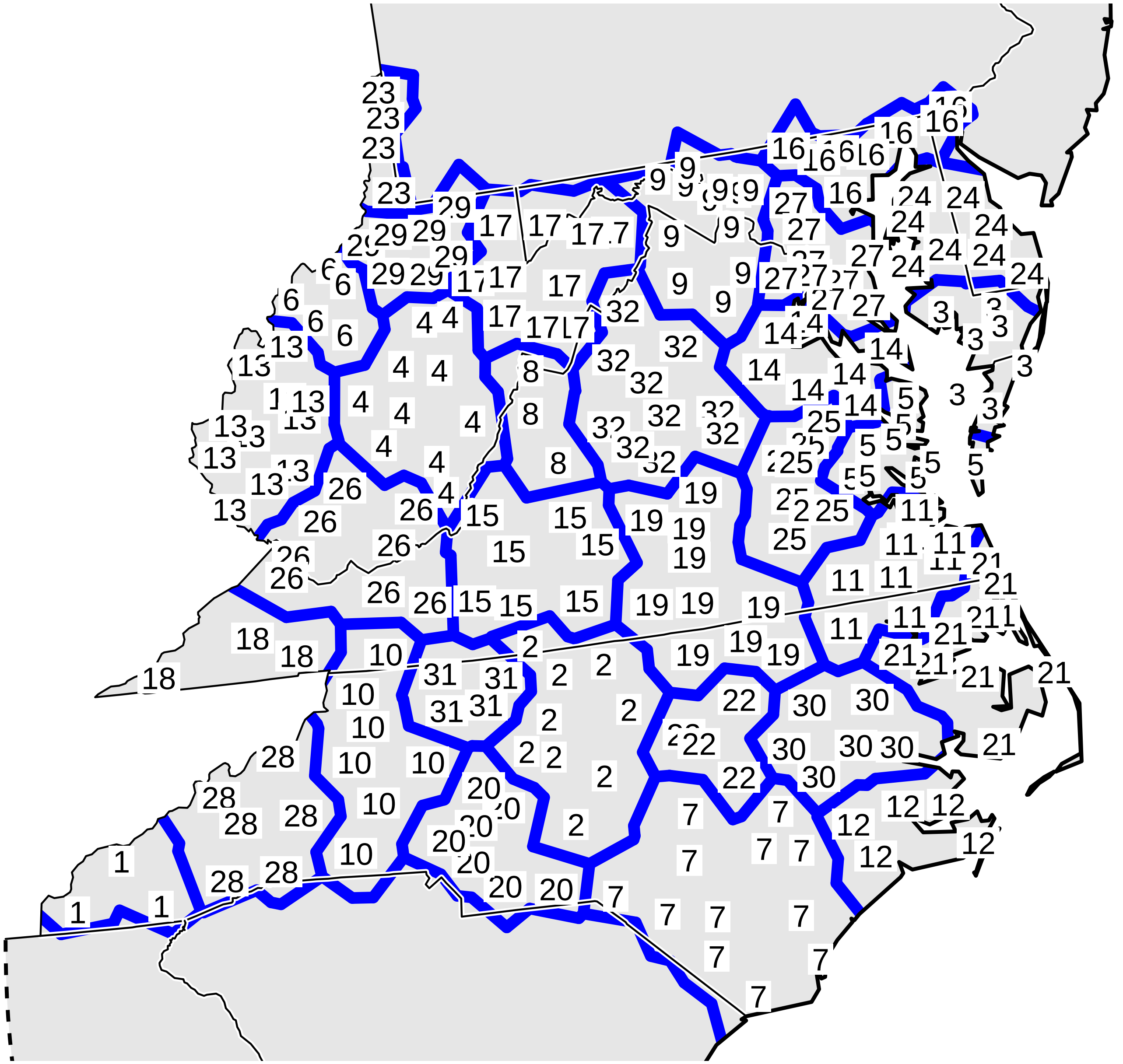
Factor Analysis

- Extract from correlation matrix those elements which reliably correlate
- Used in social science research to find common (underlying), e.g., in questionnaires
  - Check reactions to local dialect vs. standard
  - Status factor: intelligence, education, knowledgeable
  - Sympathy factor: honest, sympathetic, unpretentious
- Leading idea: examine correlations among linguistic variables, extract commonalities

Material

- Separate LAMSAS material into roughly 200 vowel pronunciations
  - first vowel in <Alabama>, last vowel in <good_morning>
- For each vowel, for each pair of sites, measure distance in vowel pronunciation
  - use LAMSAS feature chart as basis for distance
- Given that factor analysis will identify vowel occurences that function similarly (in distinguishing sites), the linguistic hypothesis is that these will reflect linguistic structure (phonemic identities, phonological processes).

Sites Grouped to Complete Matrices



Site Matrices

Per vowel we obtain a distance matrix (site $\times$ site):

	Wheeling	Winston	Raleigh	Richmond	Charlotte
Wheeling	0	41	44	45	46
Winston	41	0	16	34	36
Raleigh	44	16	0	37	38
Richmond	45	34	37	0	20
Charlotte	46	36	38	20	0

We then derive for each pair of vowels, the correlation coefficient, i.e., the degree to which they indicate the same distance between sites.

Vowel Matrix

Per vowel-pair we obtain correlation coefficient (vowel $\times$ vowel) correlations:

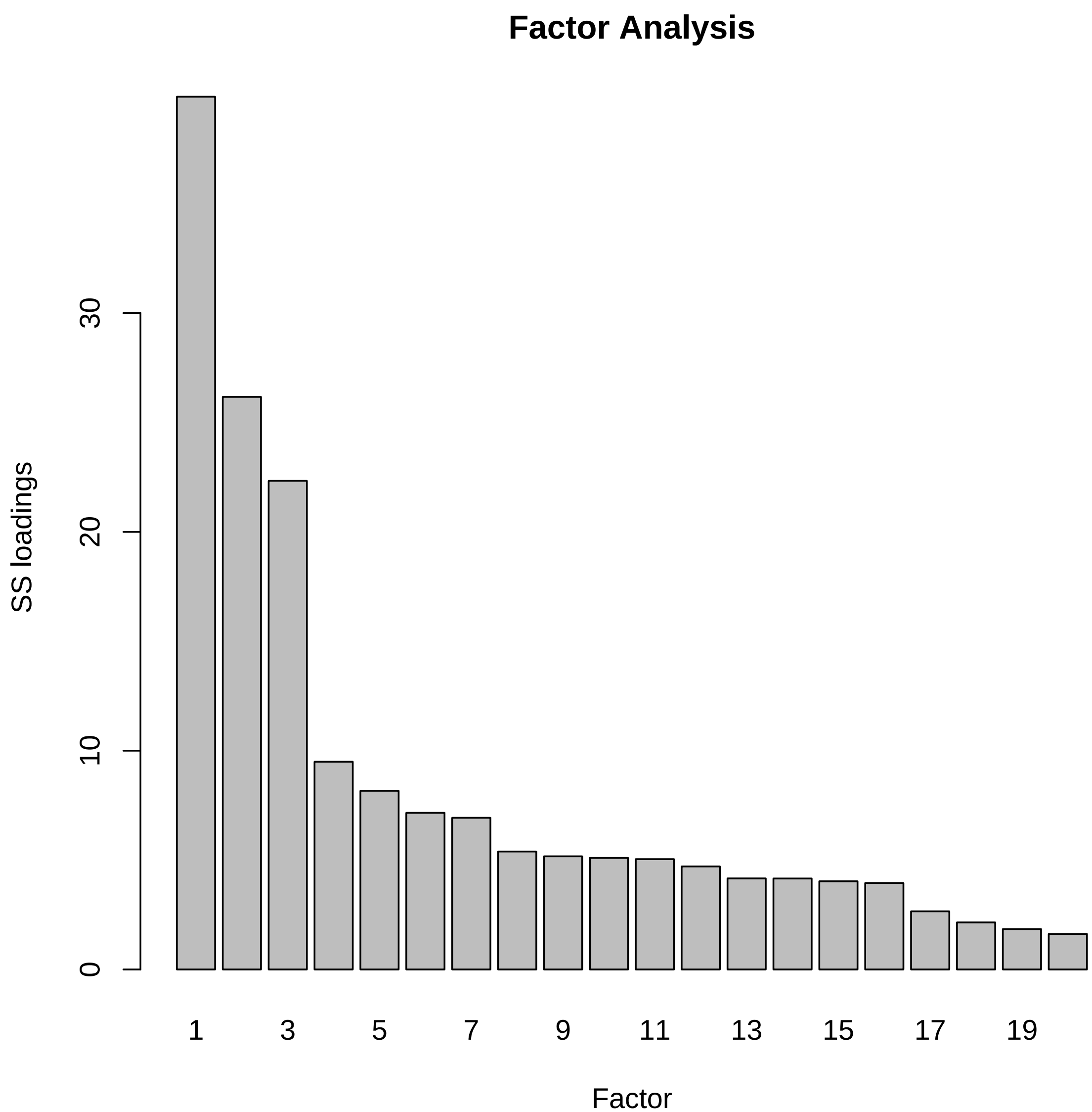
	morning1	Tuesday2	pallet2	thunderstorm2	first1
morning1	1	0.02	-0.01	0.73	0.056
Tuesday2		1	0.23	-0.03	0.02
pallet2			1	0.006	0.09
thunderstorm2				1	0.043
first1					1

This CORRELATION MATRIX is analysed for COMMON FACTORS.

We used varimax as an estimation procedure (in R): only orthogonal, no oblique rotations.

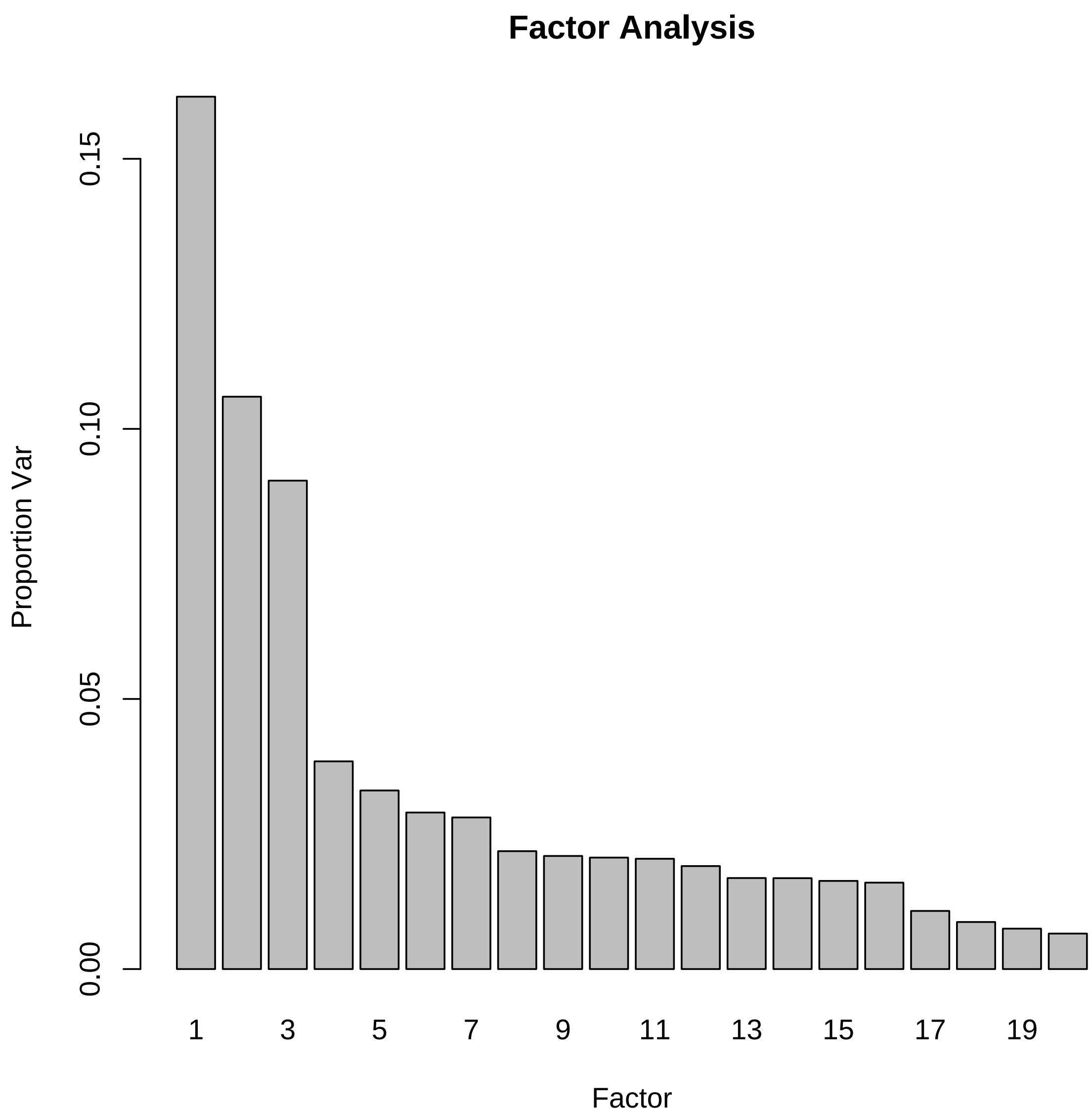
Condition: KCM/Bartlett's test of sphericity (variables are sufficiently distinct): $p < 0.001$

Loadings

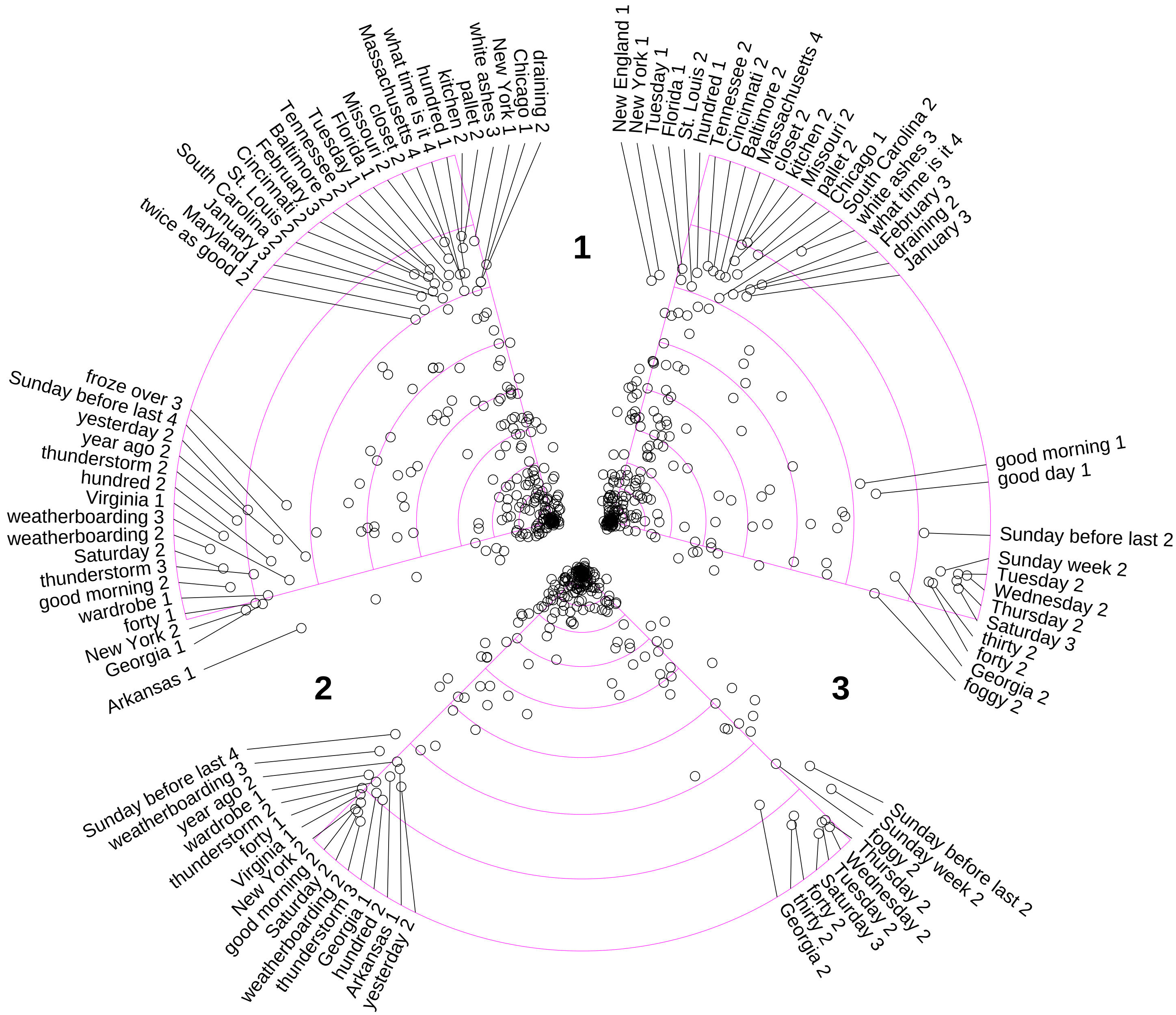


Factor Analysis
SS loadings
0
10
20
30
1
3
5
7
9
11
13
15
17
19
Factor

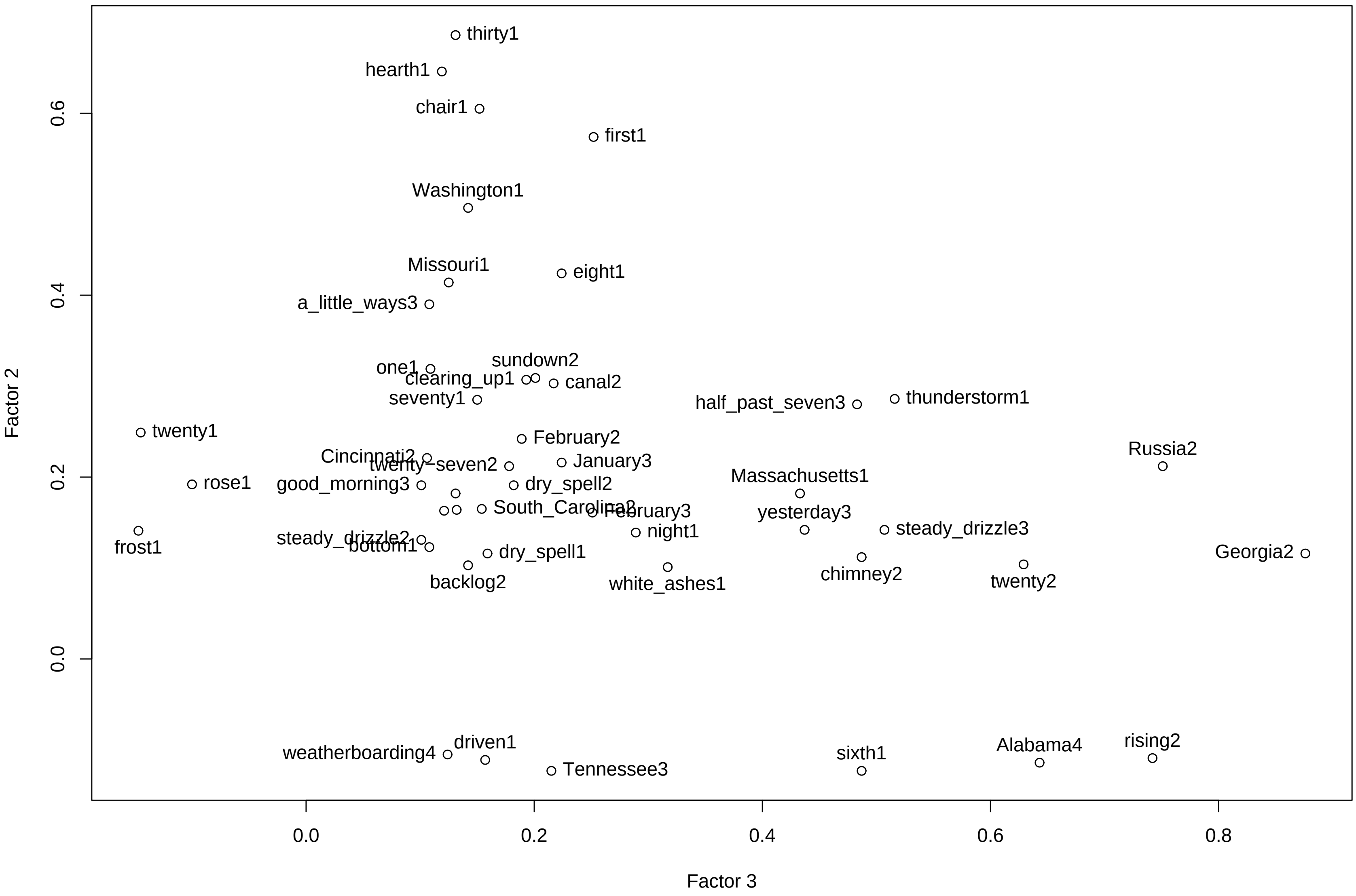
Importance of Factors



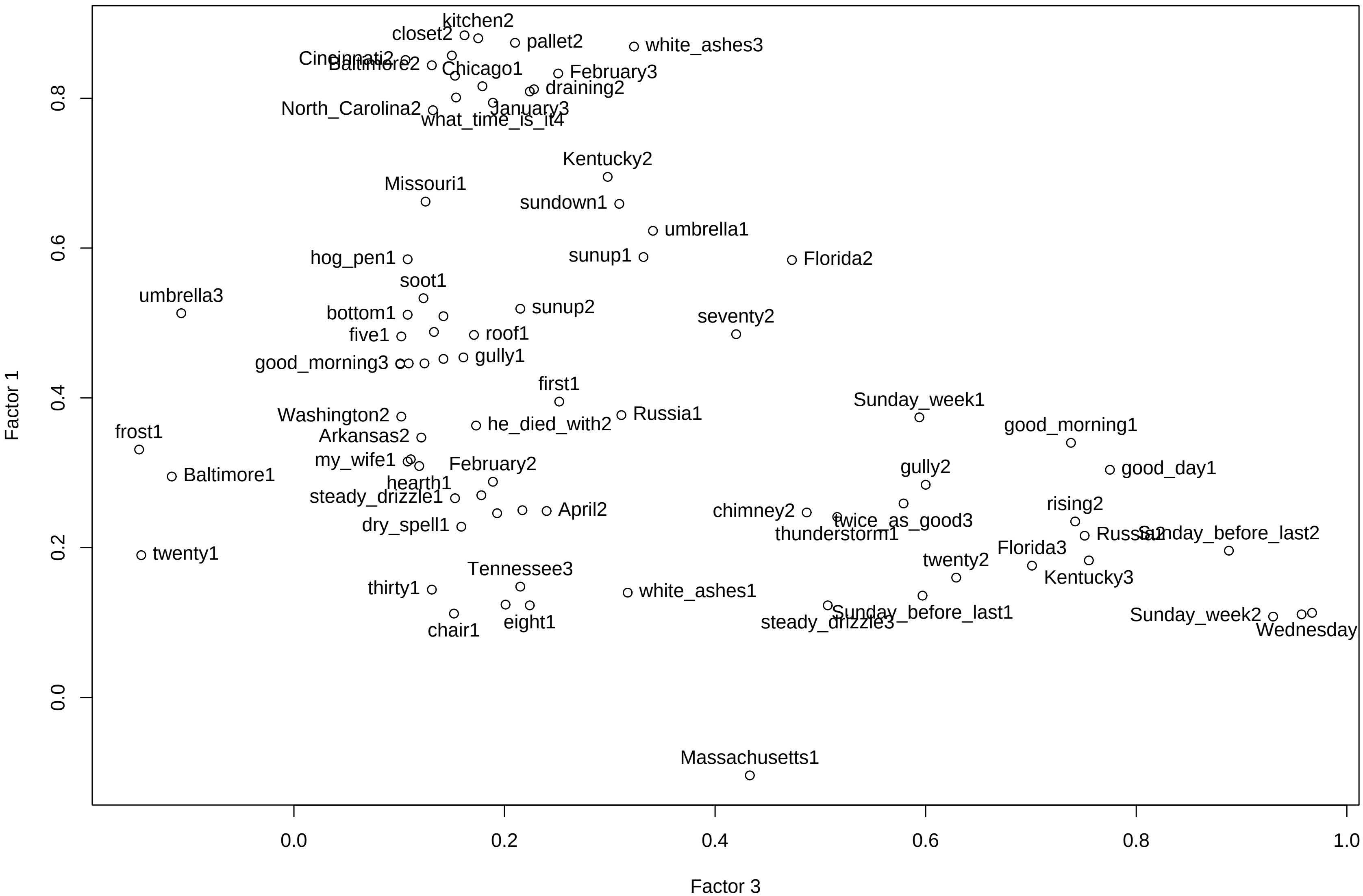
Extreme Factor Loadings



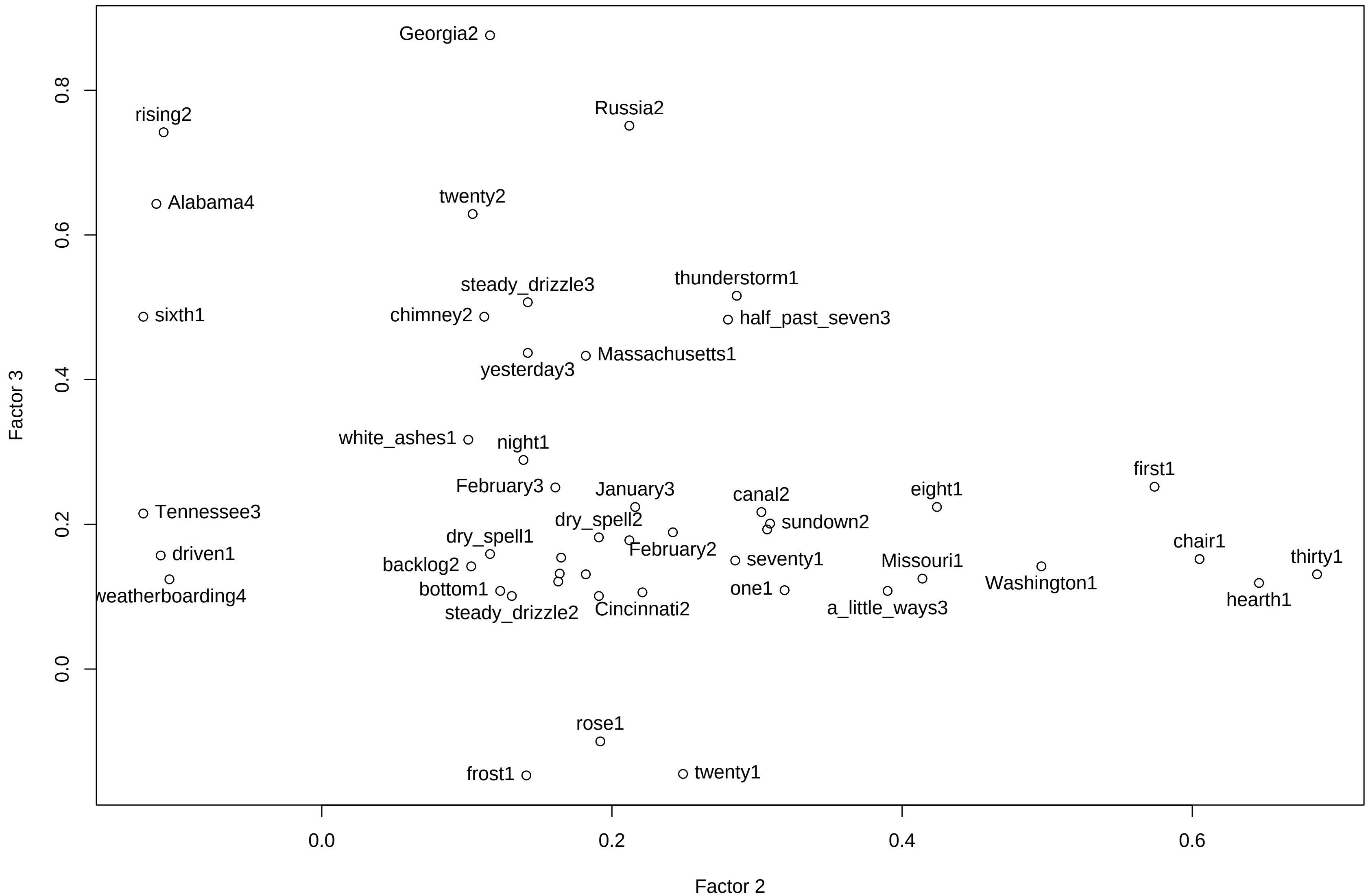
Extremes on Factors 3 & 2



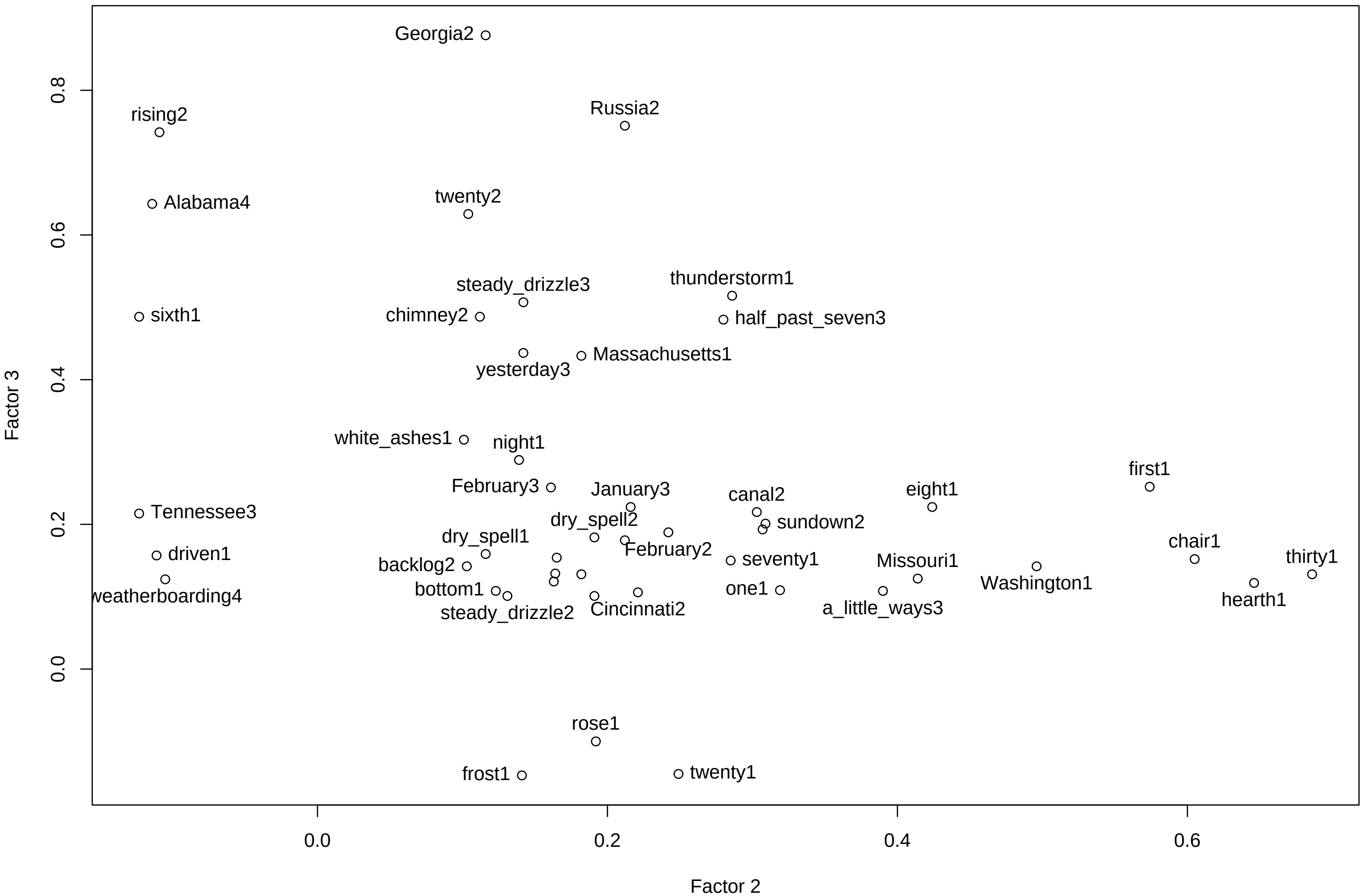
Extremes on Factors 3 & 1



Extremes on Factors 3 & 2



Extremes on Factors 3 & 2



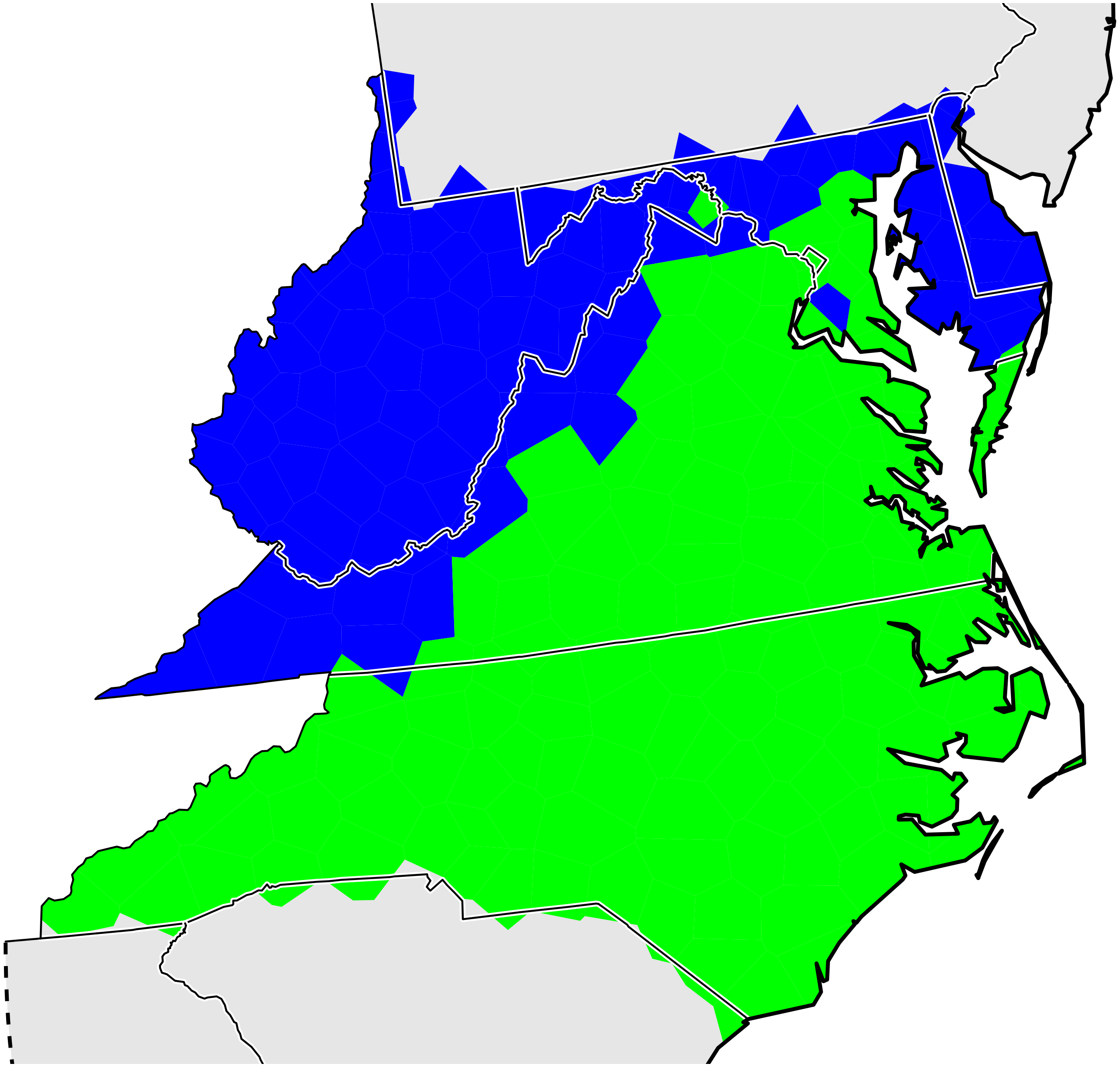
Factor 1 Loadings

closet2	0.884	kitchen2	0.880
pallet2	0.874	white_ashes3	0.869
Tennessee2	0.856	Cincinnati2	0.851
Baltimore2	0.844	Massachusetts4	0.830
Chicago1	0.816	draining2	0.812

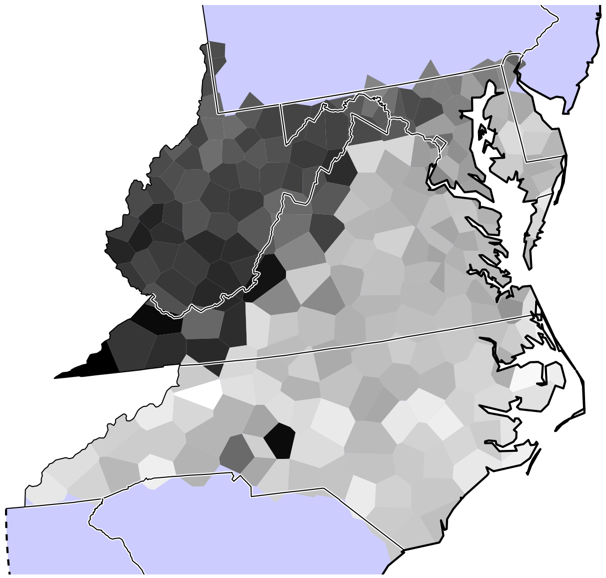
[ə] vs. [ɨ]

Florida1	0.842	[ɔ] vs. [ɑ]	St._Louis2	0.821	[u] vs. [ʉ]
hog_pen1	0.585	[ɔ] vs. [ɑ]	Tuesday1	0.796	[u] vs. [ʉ]
			Missouri2	0.857	[ʉᵊ] vs. [ʉᵊ̘]

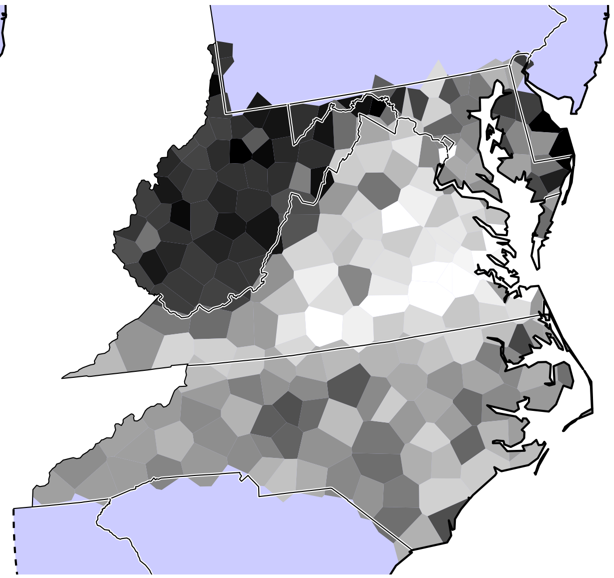
Factor 1: Geography



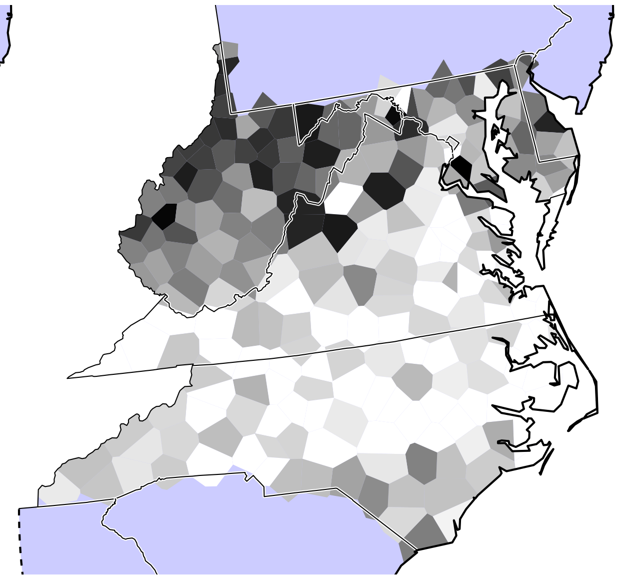
Phonological Alternations Factor 1



[ə] vs [ɨ]



[ɔ] vs [ɑ]



[u] vs. [ʉ]

Conclusion

- The first factor is sensitive to phonological alternations along the North-South division

Factor 2 Loadings

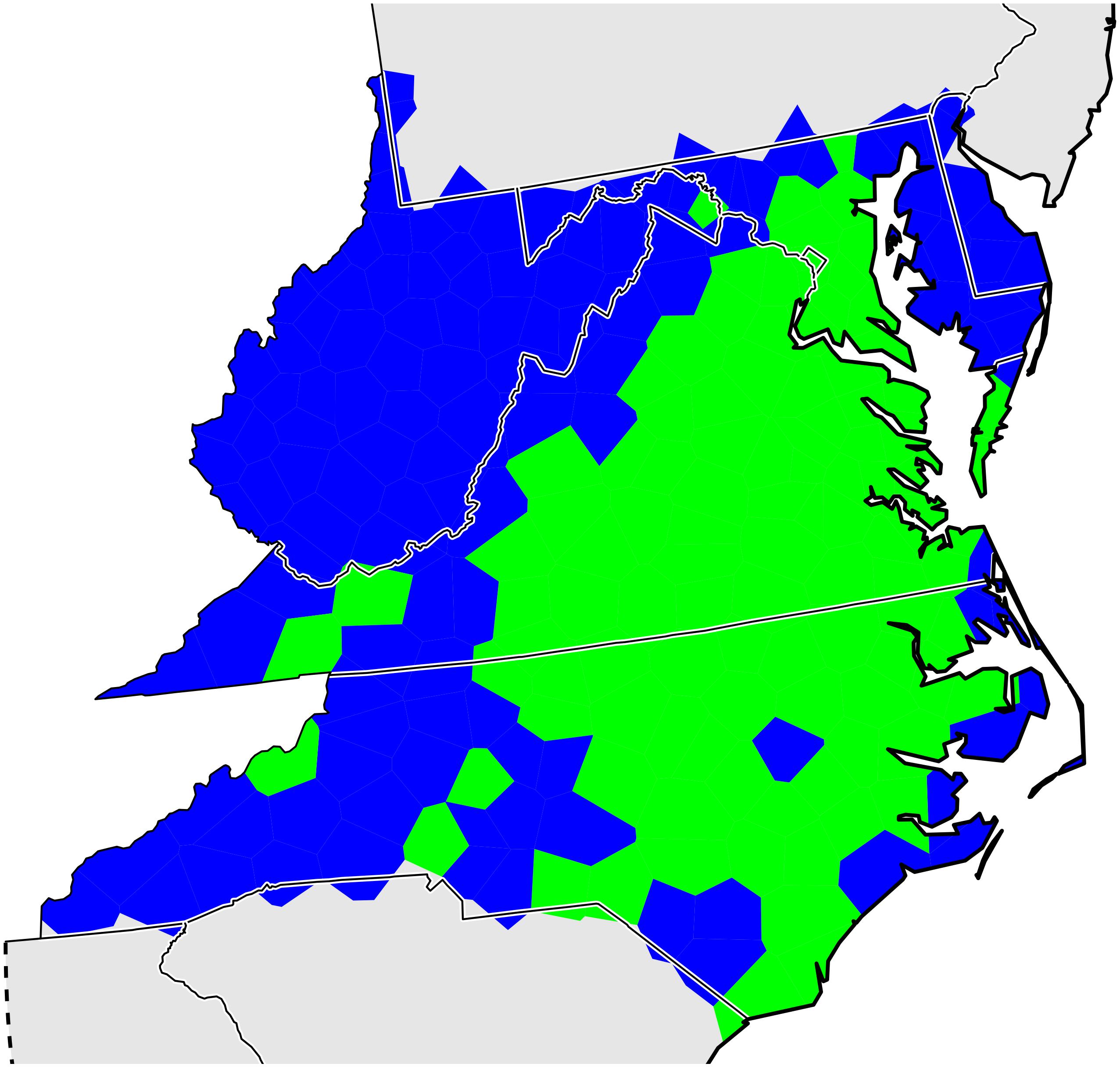
weatherboarding2	0.936	Saturday2	0.926
Virginia1	0.905		

[Vr] vs. V] (including [ɚ] vs. [ə])

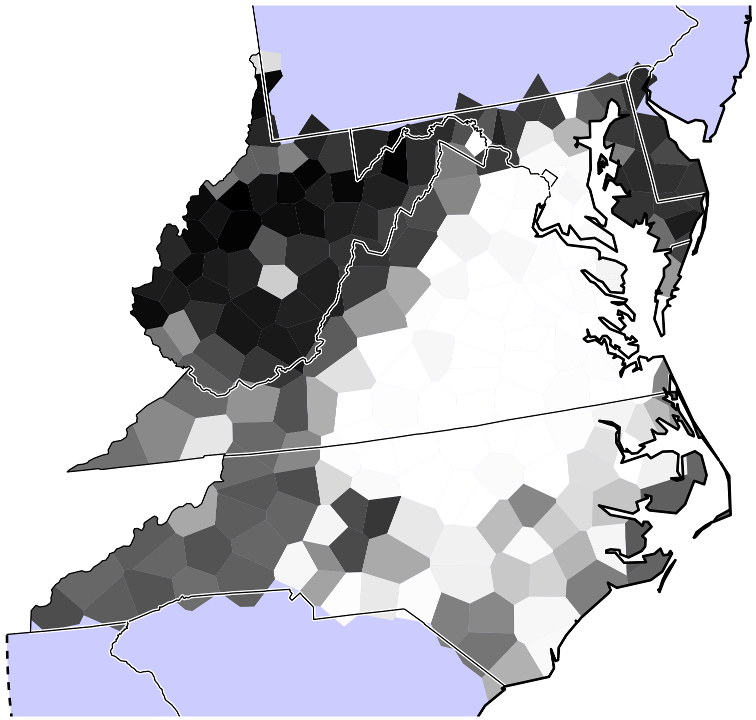
good_morning2	0.929	New_York2	0.922
forty1	0.906	thunderstorm3	0.893

[ɔə] vs. [ɔ˞ ə]

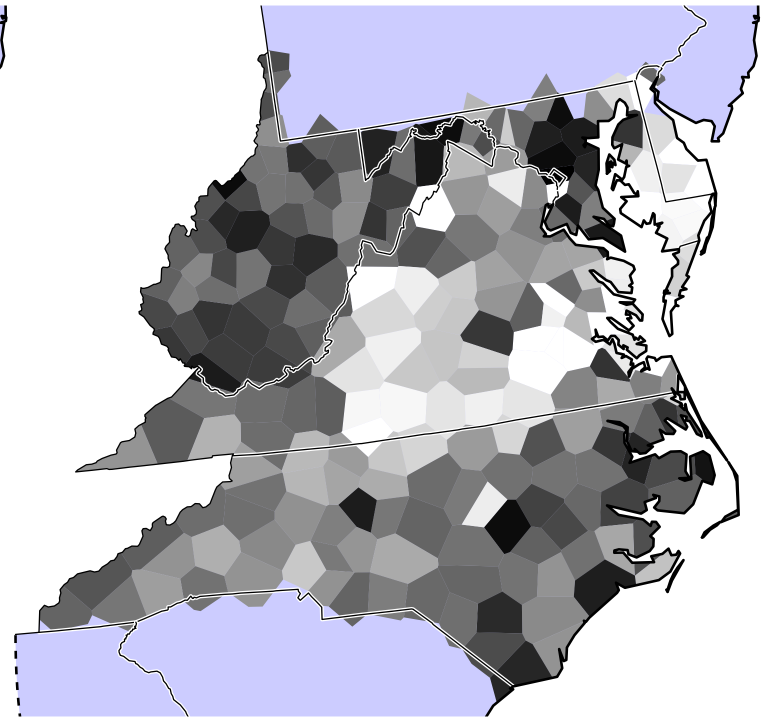
Factor 2: Geography



Phonological Alternations Factor 2



[ɚ] vs. [ə]



[ɔə] vs. [ɔ˘ ə]

- The second factor is sensitive to alternations distinguishing the Piedmont area, especially the absence of syllable final [r].
- Does [r]-lessness promote the lowering of [ɔ]?

Factor 3 Loadings

Wednesday2	0.967	Saturday3	0.961
thirty2	0.928	foggy2	0.854

[ɨ˄] vs. [ɨ]

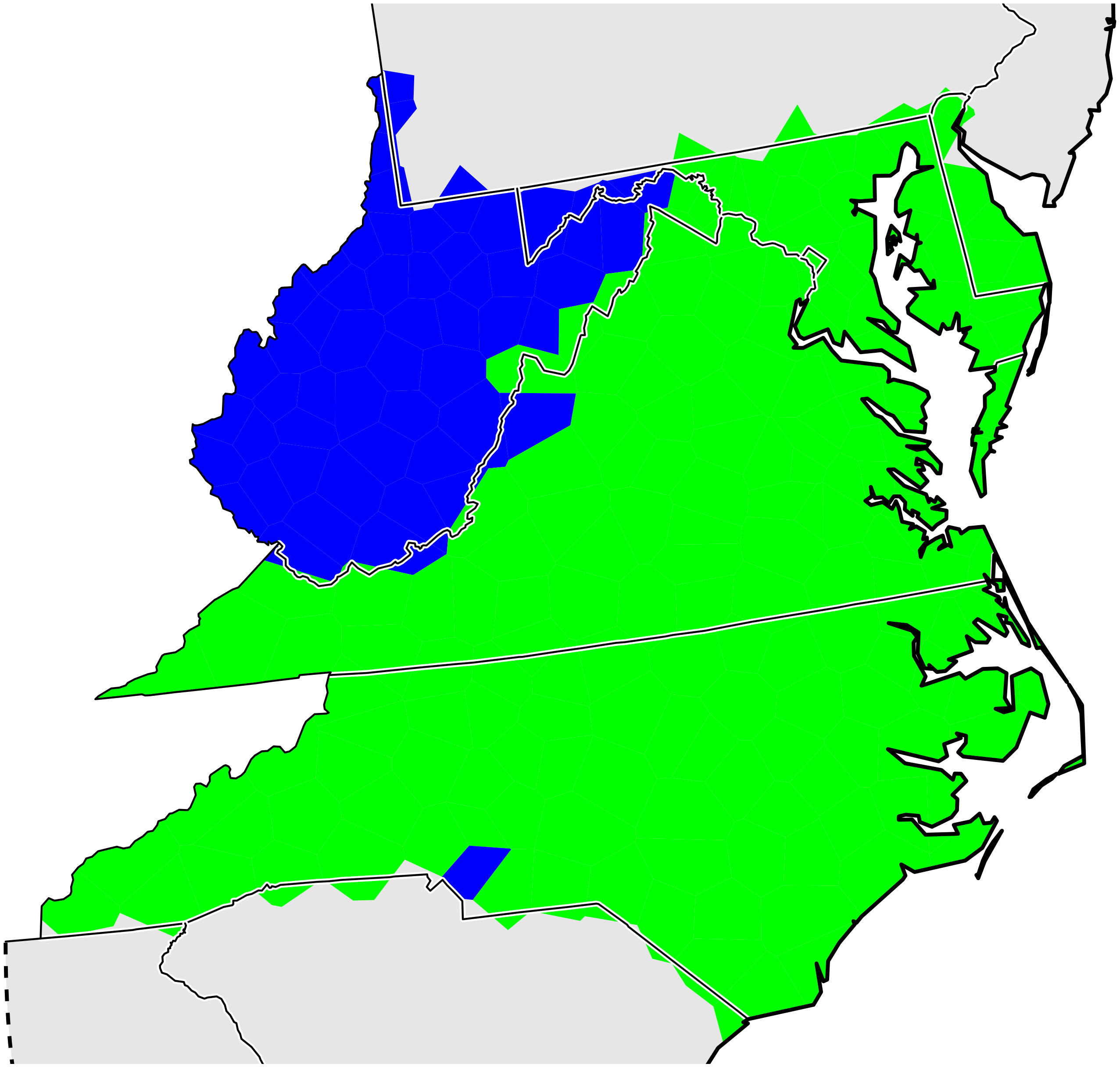
Georgia2	0.876	Tennessee1	0.766
sofa2	0.760	good_day1	0.775
Russia2	0.751	good_morning1	0.738

[ə] vs. [ɨ] (!)

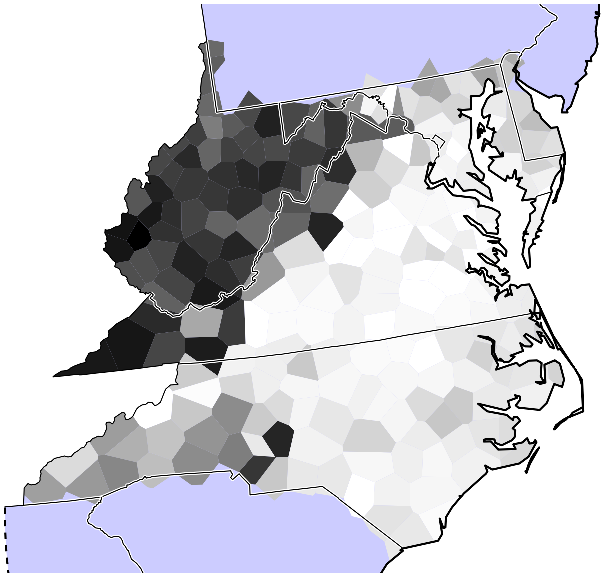
[ɛ] vs. [ɛ˄]

[ʊ] vs. [ʊ˖]

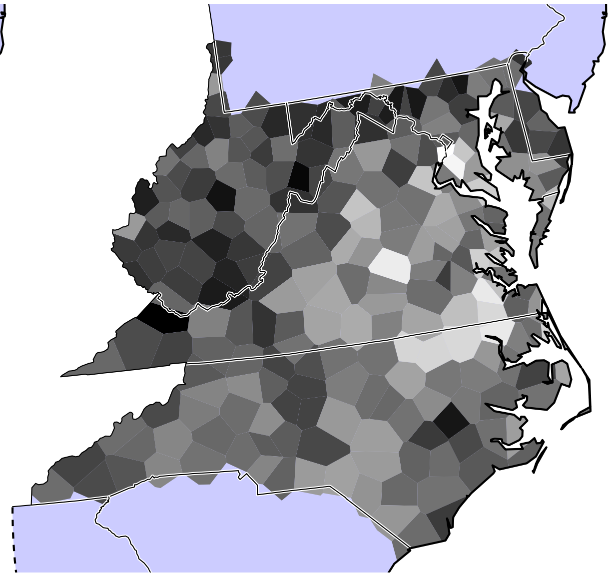
Factor 3: Geography



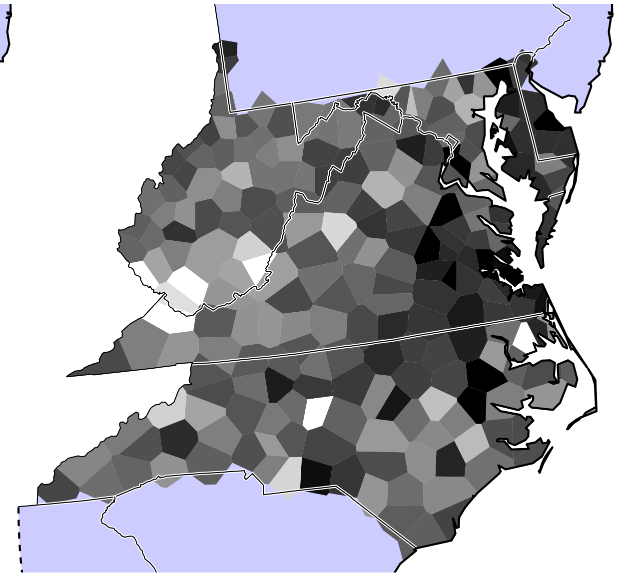
Phonological Alternations Factor 3



[ɨ˰] vs. [ɨ]



[ɛ] vs. [ɛ˰]



[ʊ] vs. [ʊ˫]

- Only the [ɨ˰] vs. [ɨ] distinction seems to pick out West Virginia as opposed to Virginia, North Carolina, Maryland, and Delaware.

Noncontrasting Vowels (in Factor Analysis)

he_died_with1
he_died_with3
New_England2
Sunday_week3
half_past_seven1
what_time_is_it1
New_Orleans2
half_past_seven2
Sunday_before_last5
steady_drizzle1
twenty-seven1
twenty-seven2
twenty-seven3
white_ashes2

April2
France1
Missouri3
attic2
backlog1
chimney1
fourteen2
eleven2
my_wife2
quilt1
seventy1
three1
thirteen2

seven2
twelve1
bureau1
ten1
bottom2
driven1
broom1
mantel1
night1
rose1
sofa1
pallet1
twenty1

kitchen1
January2
St._Louis1
second2
froze_over1
dry_spell1
froze_over2
hog_pen2
northeast2
second1
tomorrow1
January1
wardrobe2

Chicago3
Louisiana3
February1
all_at_once1
Alabama2
dry_spell2
Tennessee3
Charleston2
northwest2
a_little_ways2
Washington3
Baltimore1
bureau2

Tentative Conclusions

- Linguistic structure is exploited in dialectal distinctions. For example, phonemic distinctions are consistent across lexical items.
- Factor analysis effectively identifies linguistic structure in mass comparision
- The technique is enabled by the numeric measure of distance between segments.
- Total explained variance is low, only 36% in the first three factors. Data is noisy.
- Some factors link non-trivial linguistic variations, e.g., [ə] vs. [ɨ] on the one hand with [ɛ] vs. [ɛ̂] on the other

Future Work

- Identifying which variations to focus on (e.g., [ə] vs. [ɨ]) wrt a given factor is subjective. Can we systematize this?

- Can this technique suggest deeper linguistic relationships, e.g., different concrete alternations that are loaded for the same factor?

- Are there more general, e.g., data-mining techniques, that could be used to probe in data for which no numerical measure of difference has been established?